DOCUMENT RESUME

ED 214 994 TM 820 251

AUTHOR Eignor, Daniel R.; Douglass, James B.
TITLE A Comparison of the One-, the Modified Three-, and the Three-Parameter Item Response Theory Models in the Test Development Item Selection Process.
PUB DATE 82
NOTE 40p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (New York, March 1982); Small print in Appendix.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Comparative Analysis; *Latent Trait Theory; *Mathematical Models; Multiple Choice Tests; *Test Construction; *Test Items
IDENTIFIERS BICAL Computer Program; Estimation; LOGIST Computer Program; *One Parameter Model; Scholastic Aptitude Test; *Three Parameter Model

ABSTRACT

This paper attempts to provide some initial information about the use of a variety of item response theory (IRT) models in the item selection process; its purpose is to compare the information curves derived from the selection of items characterized by several different IRT models and their associated parameter estimation programs. These comparisons are made both visually and through the use of relative efficiency curves. Also, it is determined under what conditions (if any) items selected to meet prespecified sets of specifications in forming specially constructed tests, where the items are characterized by BICAL-based one-parameter model item parameter estimates or LOGIST-based modified three-parameter model item parameter estimates, provide as much information as items selected for the specially constructed tests based on LOGIST-based three-parameter model item parameter estimates. (Author/GK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED214994

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☒ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

A Comparison of the One-, the Modified Three-, and the Three-Parameter Item Response Theory Models in the Test Development Item Selection Process[1,2]

Daniel R. Eignor
Educational Testing Service

James B. Douglass
Opinion Research Corporation

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
D.R. Eignor
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

TM 820 251

[1] A paper presented at the annual meeting of National Council on Measurement in Education, New York, 1982.

[2] The authors gratefully acknowledge the advice provided by Frederic M. Lord and Martha L. Stocking and the technical assistance provided by Edwin O. Blew and Karen K. Zeis.

A Comparison of the One-, the Modified Three-, and the Three-Parameter Item Response Theory Models in the Test Development Item Selection Process

Daniel R. Eignor
Educational Testing Service

James B. Douglass
Opinion Research Corporation

In recent years, there has been considerable interest expressed and information written, both on a theoretical and applied level, that deals with the application of item response theory (IRT) to the test development process. On a theoretical level, based on the initial work of Birnbaum (1968), Lord (1977, 1980) and Wright and Stone (1979) (see also Wright and Douglas, 1975), have outlined methods for developing tests using the three-parameter logistic and the one-parameter logistic (Rasch) models, respectively. Lord's (1980) suggested procedure is based on the specification of a target information curve and the subsequent selection of items, based on item information, to "fill-up" the target information curve. (Item and test information are discussed in greater detail in the methodology section of this paper.) Wright and Stone (1979) discuss a procedure whereby the ability distribution of the group for which the test is intended is first specified and items are chosen so that the distribution of item difficulties matches the distribution of abilities. As pointed out by Hambleton (1979), the procedure discussed by Wright and Stone is equivalent to the one originally introduced by Birnbaum (1968) since, for the Rasch model, item information curves depend only on the difficulty parameters. In addition to this work, Hambleton (1979) and Cook and Hambleton (1978) have discussed further the advantages of using the three-parameter model in the test development process, and Cook and Hambleton (1979) have experimented with item selection procedures for "filling-up" target information curves.

The applied IRT test development work done to date has involved either the application of an IRT model to the construction of a new test or the analysis of an existing test (see Rentz and Bashaw, 1977, for distinctions between test construction and test analysis), and the bulk of this work has involved the use of the Rasch model. Two well known criterion-referenced tests, the Woodcock Reading Mastery Test (Woodcock, 1974) and the Key Math Diagnostic Arithmetic Test (Connolly, Nachtman, and Pritchett, 1971) have been constructed utilizing the Rasch model. In addition, the Rasch model has recently been used in the development of the Stanford Achievement Test (Lenke and Rentz, 1982). There have been numerous other applications of the Rasch model to test construction and analysis; Rentz and Rentz (1978) have provided a good review of this work. The three-parameter logistic model has been used for test construction purposes to build the Comprehensive Tests of Basic Skills, Form U (Yen, 1982) and has been used in a variety of test analysis contexts; see Hambleton, et al, (1978) for a review of this literature.

While theoretical work has been done articulating procedures for developing tests using the one- and three-parameter models, and a good deal of work has been done applying these methods to test construction and test analysis activities, little research has been done that compares the utility of different IRT models in the test development item selection process. This would seem to be an important avenue of research given that in the future increasingly more test development people, interested in the benefits that can be derived from IRT, will be faced with the practical decision of which IRT model to use in their work. While in most instances the nature of the data should dictate which IRT model to use (for instance, are the items amenable to guessing or do they vary greatly in their ability to discriminate), in other instances,

a look at the effects of using a variety of IRT models in the actual test development process may prove enlightening. This paper can be viewed as an attempt to provide some initial information about the use of a variety of IRT models in the item selection process; hopefully the information will assist test developers in the task of deciding on a particular IRT model for this purpose. Stated generally, the purpose of this study is to compare the information curves derived from the selection of items characterized by several different IRT models and their associated parameter estimation programs; these comparisons will be made both visually and through the use of relative efficiency curves. More specifically, the purpose is to determine under what conditions (if any) items selected to meet prespecified sets of specifications in forming specially constructed tests, where the items are characterized by BICAL-based (Wright, Mead, and Bell, 1979) one-parameter model item parameter estimates or LOGIST-based (Wood, Wingersky, and Lord, 1976) modified three-parameter model item parameter estimates, provide as much information as items selected for the specially constructed tests based on LOGIST-based three-parameter model item parameter estimates.

The outcome of this study is important for the following reason. If it is found, for instance, that there is no great loss in information and relative efficiency when items are selected using one-parameter item difficulty estimates from BICAL, compared to when items are selected using three-parameter item difficulties alone, or item difficulties and item discriminations, from LOGIST, then the simpler model and related computer program can reasonably be used for item calibration and final form item selection purposes after a pretest administration. This is important both because samples upon which pretest data are generated are usually small when the use of LOGIST and the three-parameter

model is considered, and because of cost. The BICAL program or the LOGIST program, when LOGIST is used with the modified three-parameter model, costs significantly less than when the same data is analyzed using the more complex model and LOGIST. Given the number of pretests (at least at ETS) from which final form items are chosen, this could amount to significant savings.

Method

Test Data and Samples

The data used in this study comes from two forms of the Scholastic Aptitude Test (SAT), verbal section. The SAT consists of six 30-minute sections: two verbal sections, two mathematical sections, one Test of Standard Written English (TSWE), and one experimental section, which is either made up of pretest items or a common item equating section to be used to equate the form to an existing form. The two verbal sections contain a total of 85 5-choice items, comprised of 25 antonyms, 20 analogies, 15 sentence completion, and several reading passages each of which is followed by a set of items based on the passage. Scores are reported for the verbal section (SAT-V) based on all 85 items. The common item equating section for the 85 item verbal section, used in this study as part of the overall pool of items, consists of 40 5-choice items (10 of each type). This common item equating section was administered with each of the 85 item verbal sections being used.

The 210 items from the two SAT-V forms and the common item equating section form the item pool from which items were selected to meet the prespecified distributions described in the procedure section. Each of the 210 items has parameter estimates characterizing it derived from three different IRT models and two different parameter estimation programs. These are the one-parameter

estimates of item difficulty obtained from the BICAL computer program (Wright, Mead, and Bell, 1979), modified three-parameter estimates of item difficulty obtained from the LOGIST computer program (Wood, Wingersky, and Lord, 1976), and three-parameter estimates of item difficulty, item discrimination and pseudo-guessing obtained from LOGIST. (These models and computer programs are explained in greater detail in the next two sections.) When running LOGIST with the three-parameter and modified three-parameter models, all 170 items from the two forms were run in conjuction with the 40-item equating section in one computer run (called concurrent calibration), and hence the parameter estimates were automatically on the same scale. When running BICAL with the one-parameter model, each of the 85 item forms were run with the 40-item equating section in separate computer runs. The relationship between the item difficulties from each run for the 40-item equating section was then used to place all item parameter estimates on the same scale (see Wright and Stone, 1979). The sizes of the samples used for calibration purposes, drawn from data from regular administrations of the SAT, were 2,578 candidates for one of the verbal forms and 2,565 for the other (5,143 candidates when considering concurrent LOGIST runs). The same samples were used with all three models and parameter estimation programs.

Measurement Models

The three models selected for use in this study were the one-parameter, or Rasch, the modified three-parameter, and the three-parameter logistic item response models. Each of these models specifies a particular mathematical relationship between person ability and the probability of a correct response to a given item. The three-parameter model can be written

$$P_i(\theta) = c_i + (1-c_i)\left(1+e^{-1.7a_i(\theta-b_i)}\right)^{-1}, \qquad (1)$$

where $P_i(\theta)$ is the probability of a correct response to item i by a person of ability θ,

c_i is the probability of examinees of extremely low ability giving a correct response to item i, i.e. the lower asymptote of the IRF,

b_i is the point on the ability scale where the slope of the IRF is maximum and is usually called the item's difficulty

a_i is proportional to the slope of P_i at $\theta = b_i$ and is called the item's discrimination,

and e is the base of the natural logarithms and is approximately equal to 2.718.

If the above model is simplified by fixing lower asymptotes at zero and by substituting a, a common value for the item discriminations in a test, for a_i, the one-parameter model is obtained

$$P_i(\theta) = \left(1+e^{-1.7a\,(\theta-b_i)}\right)^{-1} . \qquad (2)$$

Another version of the one-parameter logistic model was developed by Rasch (1960). This version is usually written with the constant 1.7 a incorporated into the θ (b) scale and can be written

$$P_i(\theta') = \left(1+e^{-(\theta'-b_i')}\right)^{-1} . \qquad (3)$$

The BICAL computer program uses the Rasch formulation (3), while the LOGIST computer program uses formula (2) in the estimation process when used with the Rasch model.

When running LOGIST with the modified three-parameter model, the following was done. When dealing with the one-parameter model, the lower asymptote, c_i, evident in the three-parameter model, is set to zero. This may, however, be an unrealistic way to characterize the data under study, and for this reason, when the modified three-parameter LOGIST runs were made, the c_i were set at a common value of .1480. The a_i were also set at a common value of .7880. These two values were selected on the basis of past experience with data of this type (Petersen, Cook, and Stocking, 1981). The model may be written as

$$P_i(\theta) = c + (1-c)\left(1+e^{-1.7a\ (\theta-b_i)}\right)^{-1}, \qquad (4)$$

and the model can be viewed as either a simplified (Cowell, 1981) or a modified three-parameter model.

Estimation of Model Parameters

Although both the LOGIST and BICAL computer program can be used to estimate person ability and item difficulty parameters based on the one-parameter item-response model, there can be considerable differences in the resulting parameter estimates. Further, these parameter estimates will be different

from the person ability and item difficulty parameters generated from the three-parameter model using LOGIST. Part of the reason for these differences is the manner in which the two computer programs deal with scale indeterminacy. The BICAL computer program solves the indeterminacy problem with the Rasch scale by setting the mean of the item difficulties in a test to zero while LOGIST fixes the scale by setting the average person ability to zero and the standard deviation of ability to one. When this is coupled with the important fact that differing numbers of parameters are estimated for the different models, the end result is that the person ability and item difficulty parameter estimates from the different programs are not only different in magnitude, but also on different scales. However, even if the parameter estimates were on the same scale, they would still differ in magnitude because the BICAL and LOGIST computer programs maximize different criterion functions. The BICAL computer program computes estimates of item difficulties, b_i, (i=1, 2, ..., n) and person abilities, θ_g, (g=1, 2, ..., N), that maximize the likelihood function

$$L = \prod_{g=1}^{N} \prod_{i=1}^{n} P_{ig}^{u_{ig}} Q_{ig}^{1-u_{ig}} , \qquad (5)$$

where n is the number of items and N is the number of examinees. Q_{ig} is equal to $(1 - P_{ig})$ and $u_{ig} = 1$ if examinee g answers item i correctly and $u_{ig} = 0$ otherwise.

Frequently in practical testing situations, examinees fail to answer items in a test. They may decide to skip over an item after reading it or they may not have enough time to reach items at the end of a test. For input to the LOGIST computer program, items consecutively omitted by an examinee at the end of an examination are usually coded "not reached" for that examinee. Items not

answered but followed by one or more items that are answered correctly or incorrectly are coded as omitted. Although some items coded omitted may not have been read and some items coded not reached may have been read, but not answered, as Lord (1980, p. 182) points out, the assumption made in coding items omitted and not reached seems the best practical assumption currently available. Following this line of reasoning the likelihood function was modified for the LOGIST computer program (Lord, 1974). The modification results in the criterion function

$$L^* = \prod_{g=1}^{N} \prod_{i=1}^{n_g} P_{ig}^{v_{ig}} Q_{ig}^{1-v_{ig}} , \qquad (6)$$

where n_g is the number of items finished by examinee g. If examinee g answered item i correctly or incorrectly, $v_{ig} = u_{ig}$. If examinee g omitted item i, then $v_{ig} = 1/A_i$, where A_i is the number of options for item i. LOGIST estimates the person and item parameters that maximize the criterion function, L*.

It can be seen by examining the criterion function that items coded not reached for an examinee are not used in estimating his/her ability. It is also apparent that the difficulty of an item toward the end of a partially-speeded test is not overestimated by assuming that the response would have been incorrect if the item had been reached. The $v_{ig} = 1/A_i$ for an omitted item i, is based on the assumption that an examinee's chance of answering omitted items correctly, were he/she to answer the item, is approximately $1/A_i$, i.e., chance level. This should be a more reasonable assumption than the assumption that the omitted item would have been answered incorrectly. Of course as the likelihood function (5) used in the BICAL program implies, omitted and not reached items are treated as incorrect in BICAL.

Information and Relative Efficiency

Once a latent-trait model is specified and examinee ability parameters estimated, the precision with which these examinee abilities are estimated is of interest. When the maximum likelihood estimate of ability is obtained, the precision of the ability estimate can be conveniently expressed in terms of the information function, also referred to as the information curve. Birnbaum (1968) defined information as a quantity inversely proportional to the squared length of the confidence interval around an examinee's ability. Thus, when information at an ability level is high, we have narrow confidence bands around our estimates. If information is low, we have wider confidence bands.

In mathematical terms, Birnbaum's information curve is given by

$$I_y(\theta) = \frac{\left(\sum_{i=1}^{n} w_i P_i'\right)^2}{\sum_{i=1}^{n} w_i^2 P_i Q_i} \quad . \tag{7}$$

In the expression above $I_y(\theta)$ is the amount of information at ability level θ provided by the scoring formula y, where

$$y = \sum_{i=1}^{n} w_i u_i \quad . \tag{8}$$

The variable u_i takes on values 0 or 1 depending on whether or not item is answered correctly; P_i is the probability of a correct answer to item by an examinee with ability level θ; Q_i is equal to $1-P_i$; P_i' is the slope of the

item characteristic curve at ability level θ; and the item scoring weights are w_i, i=1, 2, ..., n.

In many applied situations, such as the test development application under study in this paper, unit scoring weights are used with the items. Equation (8) then simplifies to the usual number-right score x, where

$$x = \sum_{i=1}^{n} u_i \quad . \tag{9}$$

The information curve for number-right score is then given by

$$I_x(\theta) = \frac{\left(\sum_{i=1}^{n} P'_i\right)^2}{\sum_{i=1}^{n} P_i Q_i} \quad . \tag{10}$$

Information curves can be plotted for values of θ along the ability continuum, and the height of the information curve provides data as to where measurement was most precise. One has to be careful in doing this however in that the height of the information curve can be misleading because of its dependence on the metric of the ability scale (Lord, 1975). Hence, if information curves are to be compared in any meaningful way, the ability metric must be exactly the same for each curve.

In a variety of test development applications, one may be interested in comparing the relative efficiency with which two different tests measure the same ability at various points on the ability scale. This can be determined by calculating the ratio of the values of the information curves for the two tests at each ability level. Whereas the shapes of information curves depend on the metric chosen for measuring ability, efficiency curves do not, and therefore they are particularly useful. For each distribution in this study, one of the

specially constructed tests, where the items were chosen using three-parameter item difficulty estimates generated by LOGIST, always served as the baseline test for relative efficiency calculations and comparisons.

Procedure

In prespecifying distributions from which specially constructed tests were to be built, using parameter estimates derived from the use of (1) BICAL and the one-parameter model, (2) LOGIST and the modified three-parameter model, and (3) LOGIST and the three-parameter model, a choice had to be made both as to whether to specify these distributions as target information curves or as distributions of item parameters having specific characteristics, and as to how many items should constitute these tests. Because the methodology for specifying target information curves has not been completely developed, this approach was not considered further. Since test developers are usually inclined to build tests to specifications on the item parameters (for instance, delta distributions), using these sorts of distributions was given serious consideration. Since the one-parameter model was being used, the prespecified distributions had to involve the item difficulties. An immediate problem came to light when considering the use of item difficulty distributions; the item difficulties were not on the same scale across models and related computer programs, and the (non-linear) transformations (see Yen, 1981) necessary to put them on the same scale are unknown. Hence a decision was made to reference the tests being constructed to the distribution of abilities of the candidates in the calibration sample within each model-computer run. At the same time, a decision was made to have the specially constructed tests be made up of 30 items; with a total pool of 210 items, sufficient numbers of items should exist at various points on the ability (and difficulty) metric to make up 30-item tests. For example, one of the 30 item tests was constructed to be peaked at 70% below on the ability distribution and to have difficulties approximately

normally distributed from 50% below on the ability distribution to 90% below on the ability distribution. For the one-parameter model and BICAL-based difficulty and ability estimates, the difficulty values corresponding to ability values at 50, 70, and 90% below on the ability distribution were located and then used to build a normally distributed 30 item test peaked at the difficulty value corresponding to the ability estimate at 70% below on the ability distribution. This was repeated independently using the modified three-parameter model and LOGIST based difficulty and ability parameter estimates, and then done again independently, using two different procedures with the three-parameter model and LOGIST-based parameter estimates. In the first procedure only difficulty parameter estimates corresponding to ability parameter estimates were used to build the 30 item tests. In the second procedure, once the difficulty parameter estimates corresponding to the ability parameter estimates were located, item discrimination estimates were used in conjunction with the item difficulties to build the 30 item tests.

This procedure of referencing to the ability distribution on a within-model basis circumvents (albeit in a rough fashion) the problem that exists when one wants to choose items with comparable difficulties across models; namely that the parameter estimates are not on the same scale and no convenient transformation exists for getting them on the same scale. By referencing the choice of items, based on item difficulty estimates, to the ability distribution on a within-model basis, a transformation to get difficulties on the same scale is not necessary.

Once the range of the ability distribution for which the specially constructed test is to be designed to measure is specified, variations in the distributions of item difficulties over that range (flat or uniform, symmetric, and normal) can be studied. In addition, variations in the item difficulties

can also be introduced by looking at the degree of speededness of the items (relative to the section they appeared for item calibration purposes). In this study, items were considered speeded if the percentage of candidates reaching the item is less than 95. 43 items from the total pool of 210 are considered speeded by this criterion. Table 1 presents the characteristics of each of the five prespecified distributions to be used for item selection purposes. Each of these distributions is described verbally and pictorally, for each of the models. The numbers written in each of the intervals of the item difficulty distributions are the numbers of items that must be selected from that interval in building the 30-item tests to the overall prespecified distribution. The necessary number of items in any interval were chosen randomly from the total number of items in the 210 item pool falling in that interval, except when three-parameter item discriminations were also used. Then the necessary number of items in an interval were chosen to be those that had the highest discrimation parameter estimates in that interval. Tables A through C in the Appendix give the distributions of abilities and corresponding distributions of item difficulties, with related item numbers for all 210 items, for each of the models being studied.

Only certain of the distributions in Table I, distributions one and four, were chosen to study variations in the degree of speededness of the items. Further, only the one-parameter model and BICAL-based item difficulty estimates were used in studying the effects of speededness because degree of speededness should have an effect on only the parameter estimated using BICAL (see section on estimation of model parameters). In constructing the 30 item tests involving speeded items, the speeded items were selected first in any interval and the remaining items needed to reach the total number specified were selected

Table 1

Distributions of Item Difficulties Used for Item Selection Purposes to Construct 30 Item Tests

Distribution	Description	One-Parameter Scale (BICAL)				Modified Three-Parameter Scale (LOGIST)				Three-Parameter Scale (LOGIST)			
		Upper/ Lower Ability	Percent Below	Number of Items	Range Of b's	Upper/ Lower Ability	Percent Below	Number of Items	Range Of b's	Upper/ Lower Ability	Percent Below	Number of Items	Range Of b's
1	Approximately uniform distribution of b's ranging from 50% below on the ability scale to 90% below on the ability scale.	1.29	90.0	5	1.10 - 1.29	1.19	90.6	5	1.00 - 1.19	1.29	90.2	5	1.00 - 1.29
				5	.80 - 1.09			5	.80 - .99			5	.80 - .99
				5	.60 - .79			5	.60 - .79			5	.60 - .79
				5	.40 - .59			5	.40 - .59			5	.40 - .59
				5	.20 - .39			5	.20 - .39			5	.20 - .39
		0.0	52.2	5	0 - .19	0.0	51.5	5	0 - .19	0.0	51.4	5	0 - .19
2	Approximately normal distribution of b's peaked at 70% below on the ability scale and ranging from 50% below on the ability scale to 90% below on the ability scale.	1.29	90.0	1	1.10 - 1.29	1.19	90.6	1	1.00 - 1.19	1.29	90.2	1	1.00 - 1.29
				5	.80 - 1.09			5	.80 - .99			5	.80 - .99
				9	.60 - .79			9	.60 - .79			9	.60 - .79
				9	.40 - .59			9	.40 - .59			9	.40 - .59
				5	.20 - .39			5	.20 - .39			5	.20 - .39
		0.0	52.2	1	0 - .19	0.0	51.5	1	0 - .19	0.0	51.4	1	0 - .19
3	Approximately symmetric distribution of b's peaked at 85% below on the ability scale and ranging from 75% below on the ability scale to 95% below on the ability scale.	1.79	95.1	1	1.60 - 1.79	1.59	95.0	1	1.60 - 1.69	1.69	94.9	1	1.50 - 1.69
				6	1.40 - 1.59			6	1.30 - 1.59			6	1.30 - 1.49
				8	1.20 - 1.39			8	1.10 - 1.29			8	1.10 - 1.29
				8	1.00 - 1.19			8	.90 - 1.09			8	.90 - 1.09
				6	.80 - .99			6	.70 - .89			6	.70 - .89
		.60	74.2	1	.60 - .79	.60	75.9	1	.60 - .69	.60	73.3	1	.60 - .69
4	Approximately uniform distribution of b's ranging from 25% below on the ability scale to 75% below on the ability scale.	.59	74.2	5	.40 - .59	.59	75.9	5	.40 - .59	.69	76.8	5	.50 - .69
				5	.20 - .39			5	.20 - .39			5	.30 - .49
				5	0 - .19			5	0 - .19			5	.10 - .29
				5	-.20 - -.01			5	-.20 - -.01			5	-.11 - .09
				5	-.40 - -.21			5	-.40 - -.21			5	-.30 - -.11
		-.60	26.4	5	-.60 - -.41	-.60	24.2	5	-.60 - -.41	-.60	26.5	5	-.60 - -.31
5	Approximately uniform distribution of b's ranging from 10% below on the ability scale to 50% below on the ability scale.	-.01	52.2	5	-.20 - -.01	-.01	51.5	5	-.20 - -.01	-.01	51.4	5	-.30 - -.01
				5	-.40 - -.21			5	-.40 - -.21			5	-.50 - -.31
				5	-.60 - -.41			5	-.60 - -.41			5	-.70 - -.51
				5	-.80 - -.61			5	-.80 - -.61			5	-.90 - -.71
				5	-1.00 - -.81			5	-1.00 - -.81			5	-1.10 - -.91
		-1.20	10.9	5	-1.20 - -1.01	-1.20	9.4	5	-1.20 - -1.01	-1.30	10.1	5	-1.30 - -1.11

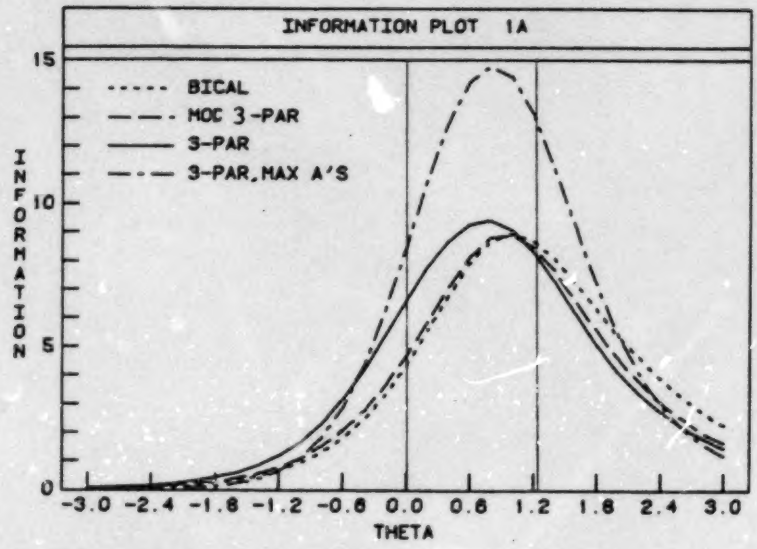
randomly. In sum, for distributions one and four, two 30 item tests were constructed using BICAL-based difficulty estimates; one 30 item test contained a certain proportion of speeded items while the other was made up entirely of unspeeded items.

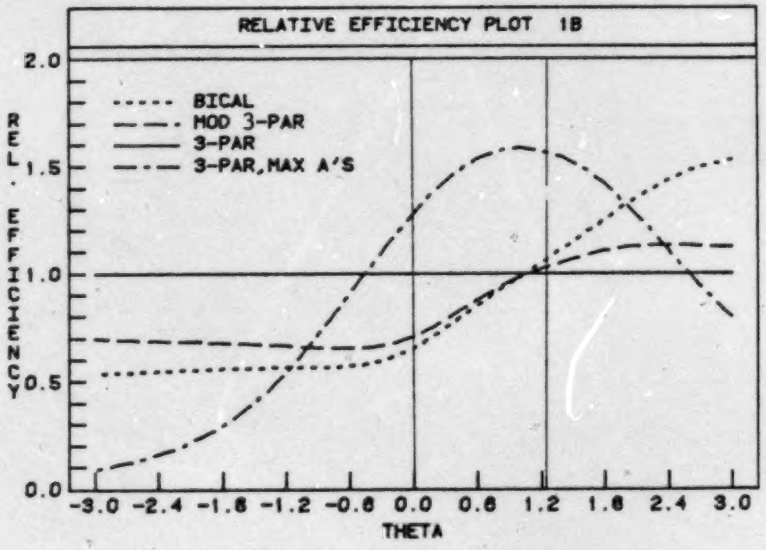
To review, for each of the five prespecified distributions of abilities and resultant item difficulties described in Table 1, 30 item tests were chosen using item parameter estimates derived from the use of 1) BICAL and the one-parameter model, 2) LOGIST and the modified three-parameter model, and 3) LOGIST and the three-parameter model. This resulted in four sets of 30 items for prespecified distributions two, three, and five (recall that two item sets were chosen using LOGIST and the three-parameter model, one set on item difficulties only and the other on difficulties and discriminations) and five sets of 30 items for prespecified distributions one and four (one Rasch set based on speeded items and one on unspeeded items); the degree of item overlap across item sets varied with the prespecified distribution. For each of the sets of 30 items for each distribution, the three-parameter LOGIST-based estimates for the items were then used to (1) construct information curves, and then (2) relative efficiency calculations were done and curves constructed. For each prespecified distribution, the 30 item test chosen using the three-parameter model item difficulties only was used as the baseline test for all relative efficiency comparisons. The relative efficiency of 30 item tests chosen using: 1) BICAL-based one-parameter item difficulty estimates; 2) LOGIST-based modified three-parameter item difficulty estimates; and, 3) LOGIST-based three-parameter item difficulty and item discrimination estimates can then be compared to the baseline test both over the range of the ability scale for which the test was constructed and in other areas of the ability scale. It should be noted again

that although the actual items for certain of the 30 items sets for each distribution were chosen based on either BICAL-based one-parameter difficulty estimates or LOGIST-based modified three-parameter difficulty estimates, the information curves for the 30 item sets were generated using the LOGIST-based three-parameter estimates for these items. As mentioned before, this is essential in that if these information curves are to be compared, the ability metric must be exactly the same. It is also important for relative efficiency comparisons, because if these curves are going to be compared in any meaningful fashion, the curves must be generated using the same ability metric. The three-parameter model was chosen as the "base model" to use for all information and relative efficiency comparisons because of the nature of the data being analyzed in this study. There is a wide range in discriminations with SAT-V data (for instance, the three-parameter LOGIST discriminations range from .31 to 1.50), and guessing on items is a common occurrence. In sum, the three-parameter model was chosen because this model most closely fits the data under study.

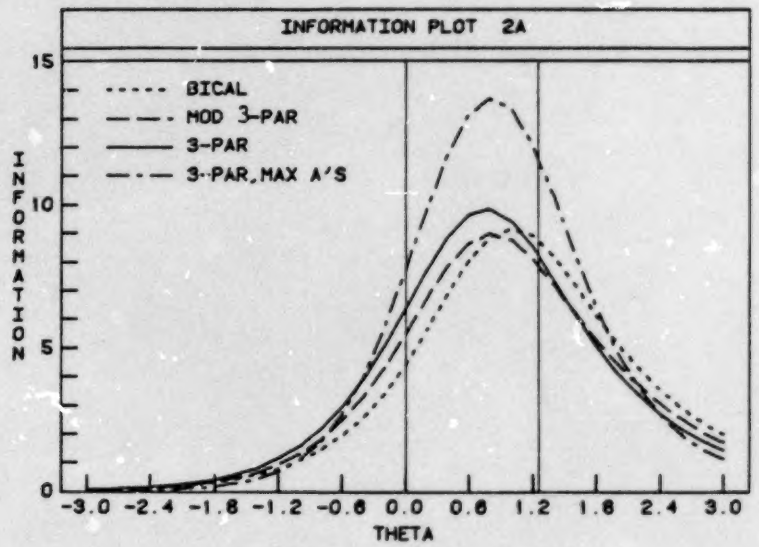
Results

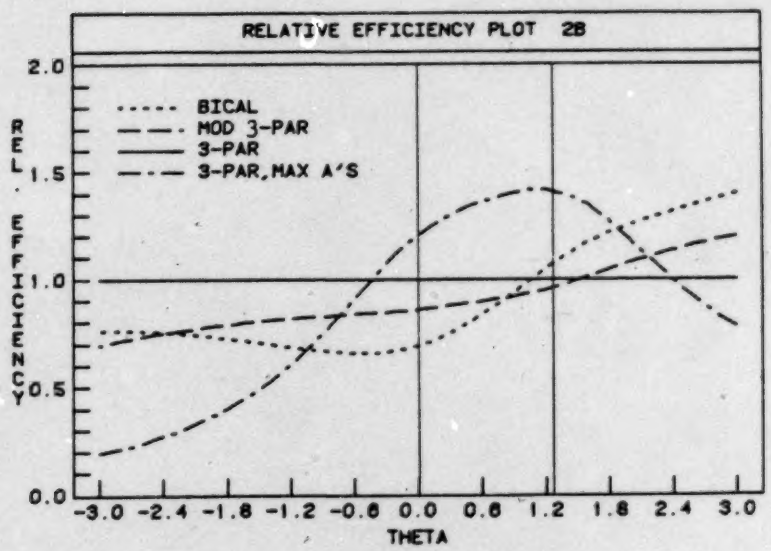
For each of the five prespecified distributions, information and relative efficiency plots were generated; the information plots are presented in Figures 1A-5A and the relative efficiency plots in Figures 1B-5B. The number identifying the plot corresponds to the number identifying the distribution in Table 1. In each of the information plots 1A-5A, three-parameter model item parameter estimates were used to generate information curves for the 30 item tests for the four item selection methods under study: 1) use of the one-parameter model and BICAL-based item difficulty estimates, 2) use of the modified three-parameter model and LOGIST-based item difficulty estimates, 3) use of the three-parameter model and LOGIST-based item difficulties alone, and 4) use of the three-parameter model and LOGIST-based item difficulties and discriminations. In the relative efficiency plots 1B-5B, the 30 item test chosen based on the three-parameter model and LOGIST-based item difficulties only was used as the baseline test for comparison purposes. It was decided that for distributions one and four, where 30 item tests were selected using speeded and unspeeded items in conjunction with the one-parameter model and BICAL, that this data be presented in separate information and relative efficiency plots. Figures 6A and 6B present the information and relative efficiency plots for distribution one and Figures 7A and 7B for distribution four. The information curves in Figures 6A and 7A are based on 30 item tests constructed using 1) Rasch difficulty estimates for speeded items, 2) Rasch difficulty estimates for unspeeded items, and 3) three-parameter model item difficulty estimates only (the baseline test for relative efficiency calculations). This choice was made because the major focus of interest lies in



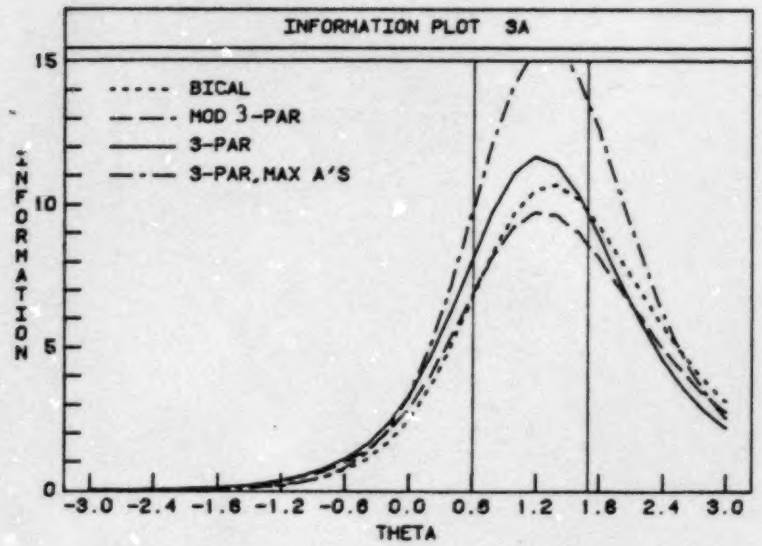


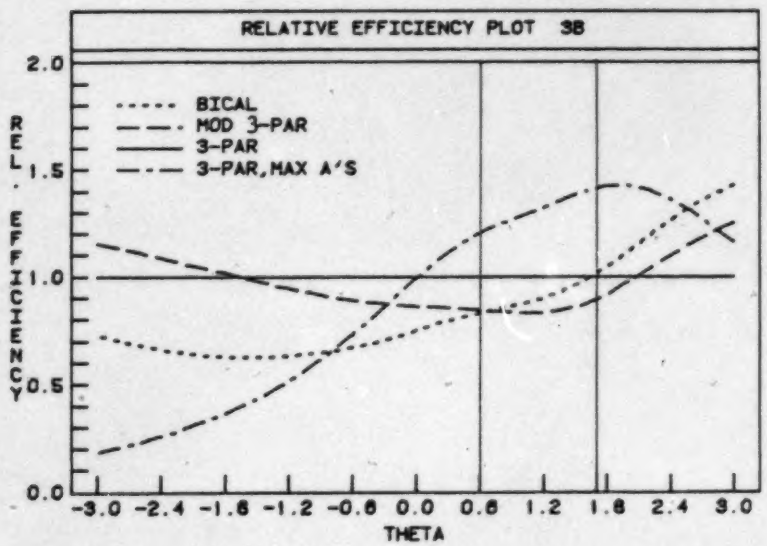
Figures 1A and 1B: Information and relative efficiency for four item selection procedures for Distribution 1 - An approximately uniform distribution of b's ranging from 50% below on the ability scale to 90% below on the ability scale.



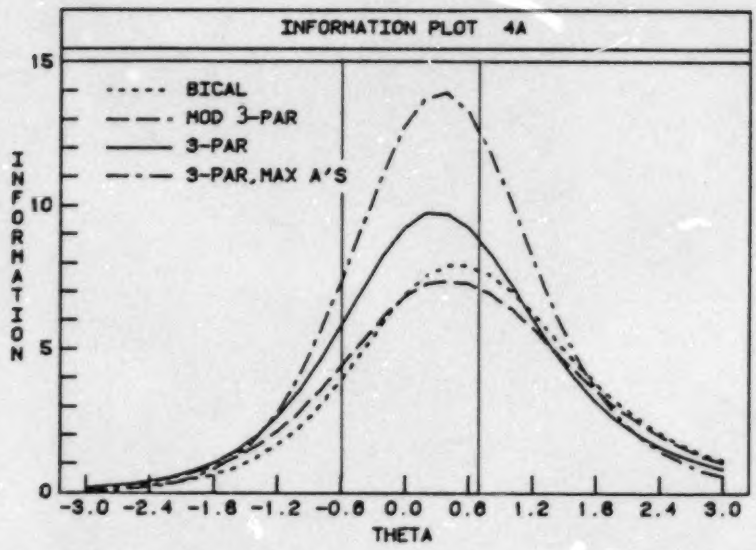


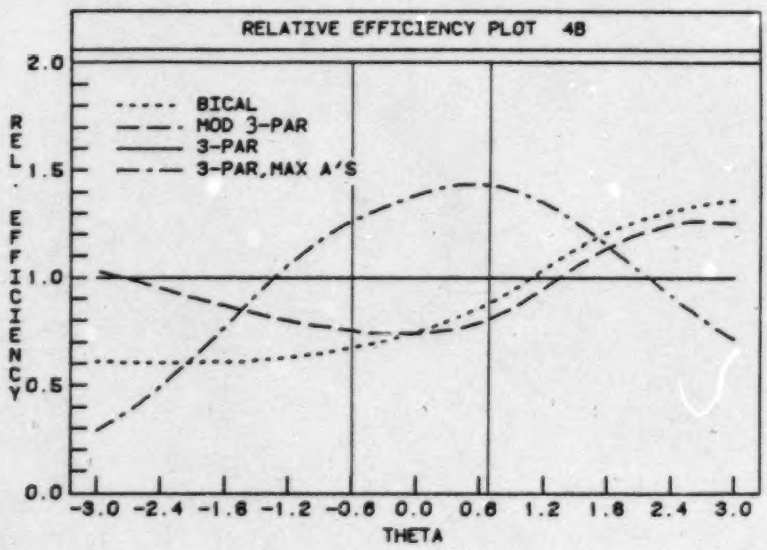
Figures 2A and 2B: Information and relative efficiency for four item selection procedures for Distribution 2 - An approximately normal distribution of b's peaked at 70% below on the ability scale and ranging from 50% below on the ability sclae to 90% below on the ability scale.



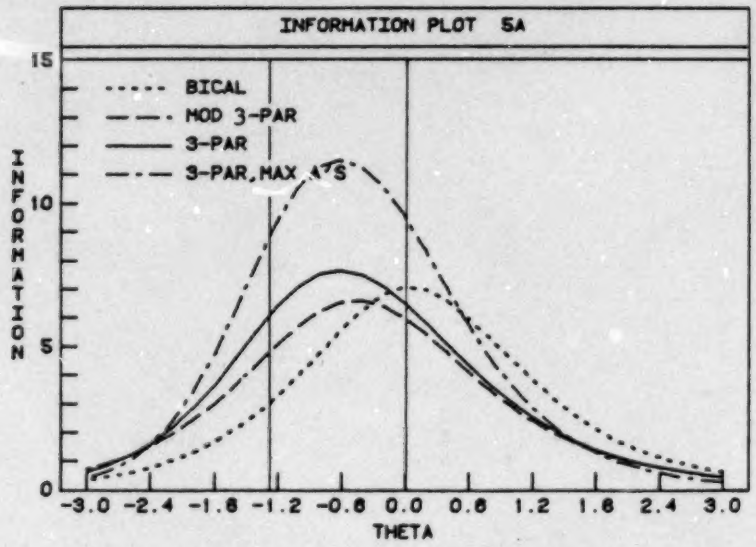


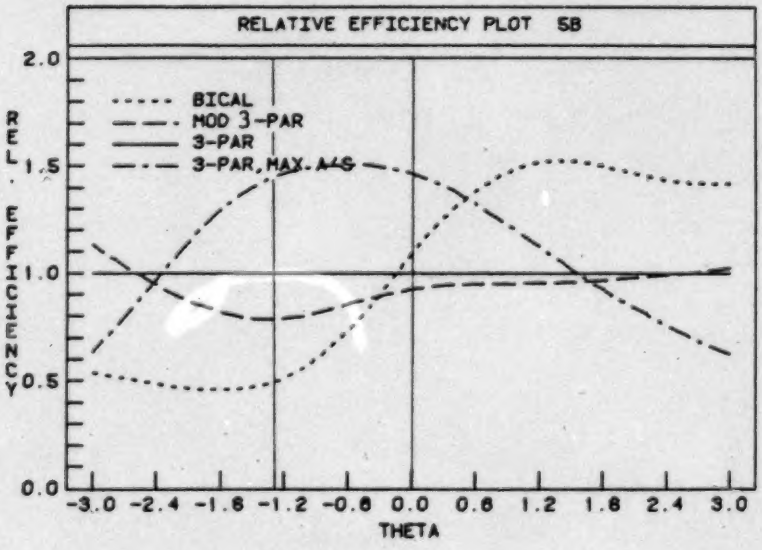
Figures 3A and 3B: Information and relative efficiency for four item selection procedures for Distribution 3 - An approximately symmetric distribution of b's peaked at 85% below on the ability scale and ranging from 75% below on the ability scale to 95% below on the ability scale.



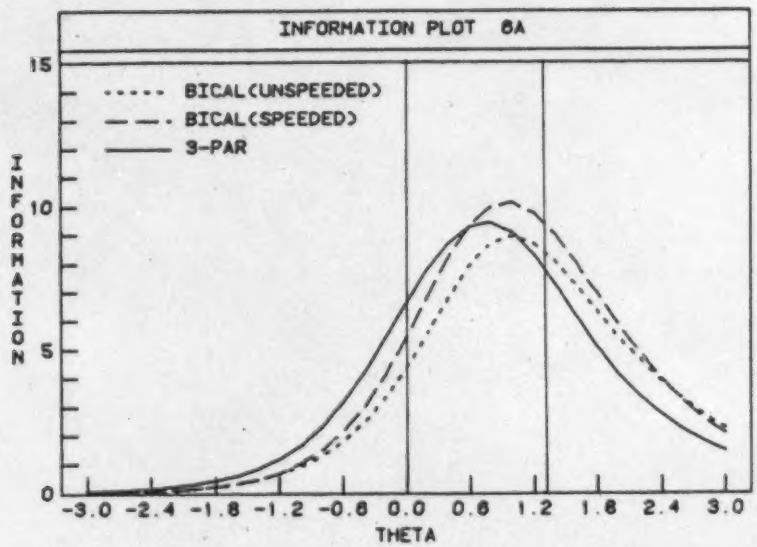


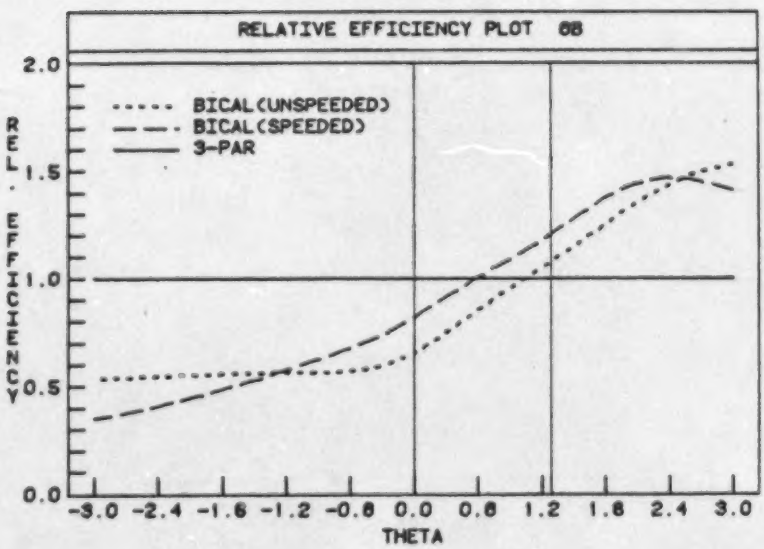
Figures 4A and 4B: Information and relative efficiency for four item selection procedures for Distribution 4 - An approximately uniform distribution of b's ranging from 25% below on the ability scale to 75% below on the ability scale.



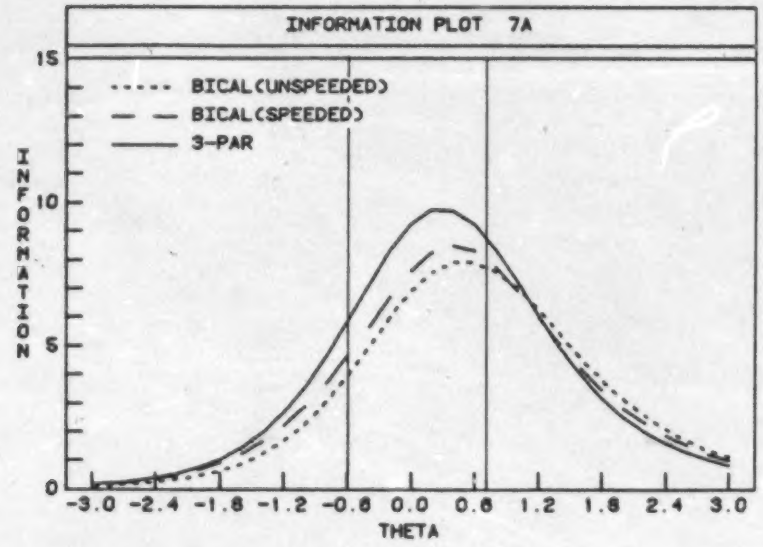


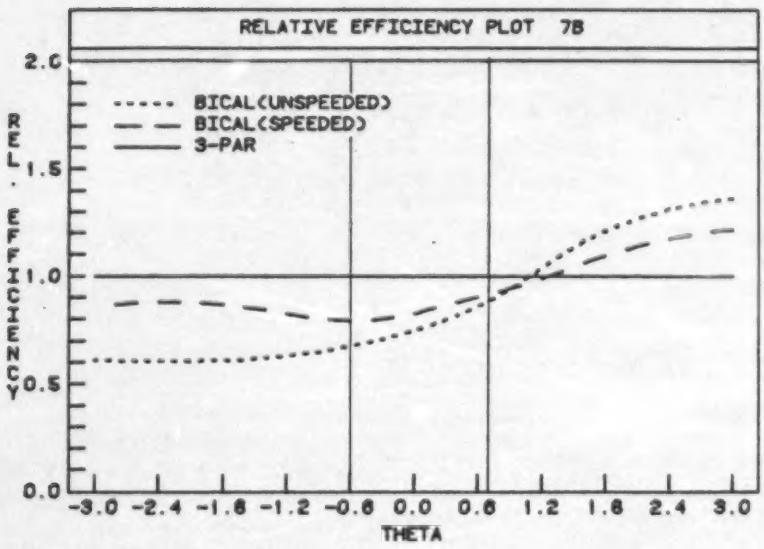
Figures 5A and 5B: Information and relative efficiency for four item selection procedures for Distribution 5 - An approximately uniform distribution of b's ranging from 10% below on the ability scale to 50% below on the ability scale.





Figures 6A and 6B: Information and relative efficiency for three item selection procedures for Distribution 1 - An approximately uniform distribution of b's ranging from 50% below on the ability scale to 90% below on the ability scale.





Figures 7A and 7B: Information and relative efficiency for three item selection procedures for Distribution 4 - An approximately uniform distribution of b's ranging from 25% below on the ability scale to 75% below on the ability scale.

comparing the information and relative efficiency curves generated for the 30 item tests built from the selection of BICAL-based item difficulties for speeded and unspeeded items.

In each of the plots, the range of the ability (theta) scale for which the specially constructed test was designed to measure has been marked off with solid vertical lines. These ranges were determined by referring to the upper and lower abilities for the three-parameter scale for each distribution in Table 1. For instance, for the information and relative plots belonging to distribution three, the solid vertical lines were drawn on the ability scale for thetas of .60 and 1.69.

The data presented in Figures 1A-5A and 1B-5B suggest the observations which follow. These observations hold across all five distributions under study.

1. Choosing items for the 30 item tests using three-parameter item difficulties and item discriminations clearly results in a test that provides more information and is more efficient in the ability range of interest than any of the other item selection procedures. The fact that this procedure is best is not an unexpected result. Had we chosen to pick items using three-parameter item information values for the ability range of interest, we would have obtained the maximum possible information curve. We chose to use only item difficulties and item discriminations for item selection purposes, rather than item information values which also take the lower asymptote into consideration, because that is the data that test development specialists at ETS are most likely to use as they begin to build tests using item response theory. Hence while it

was not unexpected that the method using item difficulties and discriminations would provide the best results, what was somewhat surprising is that none of the other three methods for selecting items provided information anywhere close in magnitude to this method. However, since this method makes use of item discrimination information, not available when items are chosen based on one-parameter or modified three-parameter item parameter estimates, the fairer comparison for these methods should be the three-parameter model item difficulties only item selection procedure.

2. Some interesting and surprising data comes to light when the information and relative efficiency plots for the 30 item tests selected using the one-parameter model and BICAL item difficulties and the modified three-parameter model and LOGIST item difficulties are compared to the 30 item tests selected using the three-parameter model item difficulties only method. For none of the five distributions, in the ability range of interest, does the item set selected using the one-parameter model or modified three-parameter model do as well as the item set chosen using the three-parameter item difficulties. Hence, if a choice be made between these three methods of item selection, the three-parameter item difficulties only method must be preferred. The differences are not, however, as great as those seen when considering the three-parameter item difficulties and discriminations method, nor are the differences consistent across the ability ranges of interest. This is the interesting data, and it is most clearly reflected in the relative efficiency plots. For each of the five distributions, in the ability range of interest,

the item set chosen using the one-parameter item difficulties is clearly less efficient at the lower end of the range of interest, but as one progresses across the range from lower to higher ability, the item set chosen using the one-parameter item difficulties becomes more efficient until, at the upper end of the range of interest, the 30 item set is either almost as efficient, as efficient, or slightly more efficient than the three-parameter item difficulties only 30 item set. A similar sort of trend occurs with the modified three-parameter model, but it is less notable. Further, for each of the distributions being considered, the 30 item set selected using one-parameter BICAL-based item difficulties always results in greater efficiency than the three-parameter model item difficulties only item set at values of theta greater than those contained in the range of interest. Again, this is also true for the 30 item sets chosen using the modified three-parameter model, but the results are clearly less striking. For thetas less than those contained in the range of interest, the item sets chosen using the one-parameter and modified three-parameter item difficulties are clearly less efficient than those chosen using the three-parameter item difficulties only method, except at very low ability levels where, for certain of the distributions, the modified three-parameter item sets become slightly more efficient.

The data presented in Figures 6A-7A and 6B-7B suggest the following observations.

1. If some of the items used for forming the 30 item tests are speeded, the use of one-parameter difficulty estimates from BICAL appears to result in a slightly more efficient test in the ability range of

interest than if the items are unspeeded. For neither distribution, however, is the 30 item test as efficient across the entire interval as that constructed using three-parameter item difficulty estimates for selection purposes. Further, the same trend of increasing efficiency as one looks across the ability range of interest from low to high is evident in the speeded 30 item test chosen using one-parameter difficulty estimates from BICAL. A possible explanation for the slightly greater efficiency for the speeded 30 item tests chosen using BICAL-based item difficulties follows. If items are speeded, the BICAL program will estimate the item difficulties to be higher (i.e., more difficult) than they, in reality, are. When choosing items for a prespecified distribution, this then results in the selection of a different set of items than those selected had the "true" item difficulties for the speeded items been used, and certainly a different set of items than those selected when the items are unspeeded. These speeded items, with "inflated" item difficulty estimates, will cause somewhat higher difficulty estimates to be picked up when the three-parameter estimates are used for information and relative efficiency curves, and since the more difficult items in this data set tend to be slightly more discriminating, will result in slightly greater efficiency in the ability range of interest.

These observations, when considered collectively, warrant two conclusions, one related to the specific underlying interest and motivation for this study, and the other more general.

1. For none of the distributions under study did the selection of items based on one-parameter or modified three-parameter item difficulties

result in a test that was as efficient over the ability range of interest as that constructed using the three-parameter model. Hence it would appear, based on the outcome of this study, that the calibration of pretest items using a simpler IRT model, and the subsequent building of final forms, based on the item parameter estimates from the simpler model, is probably not a reasonable procedure to use.

2. For data such as that used in this study, where it is more likely that the three-parameter model best fits the data, if the one-parameter model and BICAL is used to choose items to build tests to measure effectively in a particular ability range, the items selected and the resultant tests will not provide maximum information within that range of interest. Instead, the test will provide greater information and measurement efficiency at ability levels greater than those contained in the range of interest.

Discussion

The data used in this study, SAT Verbal data, is data that is probably best fit by the three-parameter model. Thus, the results which suggest that item selection procedures using the three-parameter model item difficulties and item discriminations, or item difficulties alone, provide more information and better measurement in the ability ranges of interest than item selection procedures that make use of item difficulty estimates from a simpler model are not particularly surprising. The results do, however, have implications vis a' vis the practical need to use a simpler model, and less expensive computer program, with pretest data. The interesting and surprising outcome of this study has to do with the behavior of the item sets selected

using the one-parameter model and BICAL-based item difficulty estimates. The item sets selected based on the BICAL-based item difficulty estimates clearly do not provide maximum information in the range of ability that they were selected to provide maximum information for. A partial explanation for this phenomenon may be the following. Had the data under study been "one-parameter data", then the information curves in Figures 1A-5A for the item sets selected using the one-parameter item difficulties method and the three-parameter item difficulties method would have been found to closely coincide. (If the simpler model fits the data, the more complex model will also.) However, since the data is "three-parameter data", these information curves cannot coincide, and the information curves for the item sets selected using BICAL-based item difficulty estimates are consistently shifted to the right, hence providing more information at higher ability levels, outside the ability range of interest. This shifting of information curves to the right, so that they are peaked at higher ability levels, is exactly the same phenomenon that is observed with the three-parameter model when the lower asymptote or guessing parameters for the items are increased. In an analogous fashion, because the one-parameter model does not take guessing into account, if it is used with "three-parameter data", it will result in a bias in the estimation of the item difficulties so that when these item difficulties are used to choose items to provide maximum information in a particular range of ability, in reality they provide such measurement only at higher ability levels. As mentioned above, this is probably only a partial explanation for what is happening but, to date, it is all we have been able to generate.

In summary, a good deal of information has been recently written (see Divgi, 1981) that brings to question the suitability of the Rasch model for

multiple choice items, where a more complex model is needed to explain the data. The outcome of this study, though based on only one data set and a limited number of prespecified distributions, leads one to question the use of the Rasch model in the test development item selection process when the items being selected are of the multiple-choice variety.

References

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Connolly, A. J., Nachtman, W., and Pritchett, E. M. Key match diagnostic arithmetic test. Circle Pines, MN: American Guidance Service, 1974.

Cook, L. L., and Hambleton, R. K. Application of latent trait models to the development of norm-referenced and criterion-referenced tests. Paper presented at the annual meeting of NCME, Toronto, 1978.

Cook, L. L., and Hambleton, R. K. A comparative study of item selection methods utilizing latent trait theoretic models and concepts. Paper presented at the annual meeting of NCME, San Francisco, 1979.

Cowell, W. R. Applicability of a simplified three-parameter logistic model for equating tests. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Divgi, D. R. Does the Rasch model really work? Not if you look closely. Paper presented at the annual meeting of NCME, Los Angeles, 1981.

Hambleton, R. K. Applications of latent trait theory to the development and use of criterion-referenced tests. Paper presented at the annual meeting of AERA-NCME, San Francisco, 1979.

Hambleton, R. K., Swaminathan, H., Cook, L. L., Eignor, D. R., and Gifford, J.A. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Lenke, J. M., and Rentz, R. R. The use of the Rasch model in the development of Stanford Achievement Tests. Paper presented at the annual meeting of NCME, New York, 1982.

Lord, F. M. Estimation of latent ability and item parameters when there are omitted responses. Psychometrika, 1974, 39, 247-264.

Lord, F. M. The 'ability' scale in item characteristic curve theory. Psychometrika, 1975, 44, 205-217.

Lord, F. M. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum, 1980.

Petersen, N. S., Cook, L. L., and Stocking, M. L. IRT versus conventional equating methods: A comparative study of scale stability. Paper presented at the annual meeting of AERA, Los Angeles, 1981.

Rentz, R. R., and Bashaw, W. L. The national reference scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Rentz, R. R., and Rentz, C. C. Does the Rasch model really work? A synthesis of the literature for practitioners. Princeton, N.J.: ERIC Clearinghouse on Tests, Measurement and Evaluation, Educational Testing Service, 1978.

Wood, R. L., Wingersky, M. S., and Lord, F. M. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. (RM-76-6). Princeton, N.J.: Educational Testing Service, 1976.

Woodcock, R. W. Woodcock reading mastery test. Circle Pines, MN: American Guidance Service, 1974.

Wright, B. D., and Douglas, G. A. Best test design and self-tailored testing. (Research Memorandum No. 19). Chicago, IL: University of Chicago, Department of Education, 1975.

Wright, B. D., and Stone, M. H. Best test design: Rasch measurement. Chicago, IL: MESA, 1979.

Wright, B. D., Mead, R., and Bell, S. BICAL: Calibrating items with the Rasch model. (Research Memorandum No. 23B). Chicago, IL: University of Chicago, Department of Education, 1979.

Yen, W. M. Using simulation results to choose a latent trait model. Applied Psychological Measurement, 1981, 5, 245-262.

Yen, W. M. Use of three-parameter item response theory in the development of CTBS, Forms U and TCS. Paper presented at the annual meeting of NCME, New York, 1982.

APPENDIX

TABLE A

DISTRIBUTION OF ABILITIES AND CORRESPONDING DISTRIBUTION OF ITEM DIFFICULTIES, WITH RELATED ITEM NUMBERS

1 PARAMETER: BICAL

	DISTRIBUTION OF ABILITIES		DISTRIBUTION OF DIFFICULTIES		
INTERVAL	FREQ.	PERCENT BELOW	FREQ.	UNSPEEDED ITEM NUMBERS	SPEEDED ITEM NUMBERS
3.40 - 3.49	0	99.9	0		
3.30 - 3.39	1	99.9	0		
3.20 - 3.29	6	99.7	0		
3.10 - 3.19	8	99.6	0		
3.00 - 3.09	7	99.4	0		
2.90 - 2.99	1	99.4	1	24	
2.80 - 2.89	6	99.3	2	15	170
2.70 - 2.79	7	99.1	0		
2.60 - 2.69	15	98.8	1		85
2.50 - 2.59	16	98.5	3	22,25,140	
2.40 - 2.49	10	98.3	2	139	44
2.30 - 2.39	25	97.8	2	55	169
2.20 - 2.29	17	97.5	4	14,110,185,195	
2.10 - 2.19	20	97.1	4	95,105,179	160
2.00 - 2.09	29	96.5	3	115	45,41
1.90 - 1.99	35	95.8	4	34,52,54,159	
1.80 - 1.89	32	95.1	6	23,90,100,145,194	83
1.70 - 1.79	39	94.3	1	193	
1.60 - 1.69	48	93.3	5	13,35,138,144	75
1.50 - 1.59	45	92.4	3	32,59	168
1.40 - 1.49	61	91.2	5	11,68,93,150,180	
1.30 - 1.39	60	90.0	2	69	205
1.20 - 1.29	98	88.0	7	53,89,109,143,192	84,207
1.10 - 1.19	60	86.8	3	12,94,136	
1.00 - 1.09	114	84.5	5	135,177,184,191	208
.90 - .99	96	82.5	2	137,178	
.80 - .89	115	80.2	0		
.70 - .79	133	77.5	3	33,56,103	
.60 - .69	160	74.2	7	104,120	42,80,81,166,209
.50 - .59	187	70.4	6	108,114,134,147,183,190	
.40 - .49	165	67.1	6	36,58,61,63	82,167
.30 - .39	181	63.4	3	70,88,99	
.20 - .29	166	60.0	8	51,57,66,98,199,204	210,165
.10 - .19	167	56.7	5	67,92,118,146,158	
0 - .09	220	52.2	5	8,142	43,73,74
-.10 - -.01	201	48.1	5	60,62,113,203	78
-.20 - -.11	272	42.6	2	149	79
-.30 - -.21	181	38.9	1	10	
-.40 - -.31	208	34.7	5	30,107,133	206,164
-.50 - -.41	197	30.7	3	112,182,202	
-.60 - -.51	212	26.4	7	9,21,28,72,117,157	124
-.70 - -.61	164	23.1	4	29,40,131,155	
-.80 - -.71	140	20.2	4	116,152,197	123
-.90 - -.81	105	18.1	3	65,174,187	
-1.00 - -.91	128	15.5	6	4,102,148	76,122,125
-1.10 - -1.01	126	12.9	6	49,87,189,201	77,163
-1.20 - -1.11	99	10.9	5	20,97,111,119,106	
-1.30 - -1.21	50	9.9	1	175	
-1.40 - -1.31	93	8.0	5	7,50,64,71,96	
-1.50 - -1.41	73	6.5	6	6,31,153,154,156,181	
-1.60 - -1.51	77	5.0	2	176,188	
-1.70 - -1.61	40	4.1	2	27,48	
-1.80 - -1.71	34	3.4	3	5,39,129	
-1.90 - -1.81	43	2.6	3	17,131,200	161
-2.00 - -1.91	23	2.1	2	16	
-2.10 - -2.01	28	1.5	5	46,101,132	121,162
-2.20 - -2.11	12	1.3	3	130,173,198	
-2.30 - -2.21	19	0.9	3	2,3,126	
-2.40 - -2.31	12	0.7	3	18,141,172	
-2.50 - -2.41	9	0.5	3	19,38,186	
-2.60 - -2.51	9	0.3	0		
-2.70 - -2.61	3	0.2	2	47,196	
-2.80 - -2.71	2	0.2	1	37	
-2.90 - -2.81	4	0.1	0		
-3.00 - -2.91	1	0.1	1	26	
-3.10 - -3.01	1	0.1	1	171	
-3.20 - -3.11	1	0.1	2	91,127	
-3.30 - -3.21	1	0.0	0		
-3.40 - -3.31	0	0.0	1	86	
-3.50 - -3.41	1	0.0	0		
-3.60 - -3.51	0	0.0	0		
-3.70 - -3.61	0	0.0	2	1,128	

TABLE B

DISTRIBUTION OF ABILITIES AND CORRESPONDING DISTRIBUTION OF ITEM DIFFICULTIES, WITH RELATED ITEM NUMBERS

MODIFIED 3 PARAMETER: LOGIST

	DISTRIBUTION OF ABILITIES		DISTRIBUTION OF DIFFICULTIES		
INTERVAL	FREQ.	PERCENT BELOW	FREQ.	UNSPEEDED ITEM NUMBERS	SPEEDED ITEM NUMBERS
3.40 - 3.49	0	100.0	0		
3.30 - 3.39	0	100.0	1	170	170
3.20 - 3.29	1	99.9	0		
3.10 - 3.19	2	99.9	1	24	
3.00 - 3.09	1	99.9	1	140	
2.90 - 2.99	2	99.8	3	15,22	44
2.80 - 2.89	6	99.7	0		
2.70 - 2.79	9	99.6	2	25,55	
2.60 - 2.69	6	99.4	0		
2.50 - 2.59	5	99.3	2		41,85
2.40 - 2.49	9	99.2	4	105,110,115	169
2.30 - 2.39	13	98.9	8	14,23,145,179,185,195	45,160
2.20 - 2.29	20	98.5	2	95,139	
2.10 - 2.19	19	98.1	2	52,159	
2.00 - 2.09	19	97.8	3	90,100,144	
1.90 - 1.99	29	97.2	3	34,54,194	
1.80 - 1.89	26	96.7	4	59,138,150	75
1.70 - 1.79	34	96.0	2	35,193	
1.60 - 1.69	53	95.0	4	11,32,180	83
1.50 - 1.59	39	94.3	5	13,69,143,192	205
1.40 - 1.49	55	93.2	3	68,93	168
1.30 - 1.39	66	91.9	5	89,136,184,191	207
1.20 - 1.29	64	90.6	3	53,109,135	
1.10 - 1.19	88	88.9	3	94,177	208
1.00 - 1.09	100	87.0	2	12,178	
.90 - .99	99	85.1	6	33,56,103,104,120,137	
.80 - .89	140	82.3	3	147,183	84
.70 - .79	160	79.2	4	36,63,190	42
.60 - .69	170	75.9	6	58,61,108,114,134,166	
.50 - .59	173	72.5	6	70,99,199	81,167,209
.40 - .49	208	68.5	10	51,57,66,88,118,146,158,204	80,165
.30 - .39	194	64.7	4	67,92,98	82
.20 - .29	233	60.2	4	8,142,203	74
.10 - .19	219	55.9	6	60,62,113,149	73,210
0 - .09	228	51.5	2		43,78
-.10 - -.01	253	46.5	3	10,30,112	
-.20 - -.11	248	41.7	9	9,21,117,133,182,202	79,164,206
-.30 - -.21	257	36.7	6	28,40,72,107,155,157	
-.40 - -.31	222	32.4	5	29,116,151,152,197	
-.50 - -.41	237	27.7	4	4,174,187	124
-.60 - -.51	182	24.2	2	65,148	
-.70 - -.61	176	20.8	7	20,87,97,102,189,201	123
-.80 - -.71	158	17.7	5	49,111,119	122,163
-.90 - -.81	134	15.1	10	7,50,64,96,106,153,156,175,181	77
-1.00 - -.91	114	12.9	6	6,31,71,176,188	76
-1.10 - -1.01	95	11.0	3	48,154	125
-1.20 - -1.11	83	9.4	3	39,129,131	
-1.30 - -1.21	83	7.8	6	5,16,17,27,173,200	
-1.40 - -1.31	58	6.6	4	101,130,132,141	
-1.50 - -1.41	47	5.7	3	46,126	161
-1.60 - -1.51	43	4.9	6	3,19,172,198	121,162
-1.70 - -1.61	43	4.1	4	2,18,38,186	
-1.80 - -1.71	29	3.5	1	171	
-1.90 - -1.81	27	3.0	2	47,196	
-2.00 - -1.91	27	2.4	2	37,91	
-2.10 - -2.01	19	2.1	3	26,86,127	
-2.20 - -2.11	17	1.7	0		
-2.30 - -2.21	15	1.4	1	128	
-2.40 - -2.31	13	1.2	1	1	
-2.50 - -2.41	8	1.0	0		
-2.60 - -2.51	10	0.8	0		
-2.70 - -2.61	3	0.8	0		
-2.80 - -2.71	6	0.7	0		
-2.90 - -2.81	5	0.6	0		
-3.00 - -2.91	4	0.5	0		
-3.10 - -3.01	5	0.4	0		
-3.20 - -3.11	3	0.3	0		
-3.30 - -3.21	3	0.3	0		
-3.40 - -3.31	3	0.2	0		
-3.50 - -3.41	1	0.2	0		
-3.60 - -3.51	2	0.2	0		
-3.70 - -3.61	1	0.1	0		

TABLE C

DISTRIBUTION OF ABILITIES AND CORRESPONDING DISTRIBUTION OF ITEM DIFFICULTIES, WITH RELATED ITEM NUMBERS

THREE PARAMETER: LOGIST

	DISTRIBUTION OF ABILITIES		DISTRIBUTION OF DIFFICULTIES		
INTERVAL	FREQ.	PERCENT BELOW	FREQ.	UNSPEEDED ITEM NUMBERS	SPEEDED ITEM NUMBERS
3.40 - 3.49	1	99.9	0		
3.30 - 3.39	0	99.9	0		
3.20 - 3.29	1	99.9	0		
3.10 - 3.19	2	99.9	1	15	
3.00 - 3.09	1	99.9	0		
2.90 - 2.99	4	99.8	0		
2.80 - 2.89	8	99.6	0		
2.70 - 2.79	7	99.5	3	115,140	170
2.60 - 2.69	5	99.4	1		44
2.50 - 2.59	9	99.2	3	14,24	41
2.40 - 2.49	13	99.0	1	23	
2.30 - 2.39	15	98.7	1		169
2.20 - 2.29	23	98.2	3	22,179,195	
2.10 - 2.19	20	97.8	3	55,145	85
2.00 - 2.09	30	97.3	5	25,105,144,194	45
1.90 - 1.99	38	96.5	6	35,53,54,69,110,138	
1.80 - 1.89	28	96.0	7	34,95,100,150,193,139	83
1.70 - 1.79	36	94.9	8	13,52,93,109,143,159,59	160
1.60 - 1.69	40	94.1	1	185	
1.50 - 1.59	60	92.9	2	12,135	
1.40 - 1.49	63	91.7	9	32,33,90,94,180,191	75,168,208
1.30 - 1.39	75	90.2	6	11,68,120,183,192	205
1.20 - 1.29	72	88.8	4	136,178,184	207
1.10 - 1.19	111	86.7	1	147	
1.00 - 1.09	114	84.5	3	36,56,89	
.90 - .99	110	82.3	7	60,104,134,137,177	84,166
.80 - .89	142	79.5	5	8,108,146,190	42
.70 - .79	143	76.8	5	57,58,142,199	165
.60 - .69	179	73.3	5	61,63,70,103	80
.50 - .59	171	69.9	5	51,114,204	167,209
..40 - .49	177	66.5	7	62,66,98,99,158,203	81
.30 - .39	199	62.6	5	92,118	74,82,210
.20 - .29	198	58.8	3	67,88	43
.10 - .19	190	55.1	4	10,113,149	73
0 - .09	190	51.4	2	21	206
-.10 - -.01	253	46.4	0		
-.20 - -.11	215	42.2	6	29,30,112,182,202	78
-.30 - -.21	221	37.9	7	9,28,72,117,152,155	164
-.40 - -.31	191	34.2	6	4,40,133,157,197	79
-.50 - -.41	198	30.4	4	106,107,151	124
-.60 - -.51	197	26.5	2	116,148	
-.70 - -.61	170	23.2	1	187	
-.80 - -.71	136	20.6	7	20,50,153,174,189,201	123
-.90 - -.81	143	17.8	2	97,119	
-1.00 - -.91	117	15.5	9	6,65,87,131,136,175,181	77,125
-1.10 - -1.01	100	13.5	6	16,48,102,188	122,163
-1.20 - -1.11	88	11.8	5	7,111,141,173,176	
-1.30 - -1.21	91	10.1	6	17,19,31,39,129,130	
-1.40 - -1.31	73	8.6	3	64,126	121
-1.50 - -1.41	56	7.5	5	46,71,171,200	162
-1.60 - -1.51	60	6.4	1	18	
-1.70 - -1.61	44	5.5	5	3,38,49,172	76
-1.80 - -1.71	42	4.7	3	5,101,132	
-1.90 - -1.81	33	4.1	3	47,86,96	
-2.00 - -1.91	33	3.4	5	27,91,127,186,198	
-2.10 - -2.01	22	3.0	3	1,128,134	
-2.20 - -2.11	24	2.5	0		
-2.30 - -2.21	20	2.1	2	26	161
-2.40 - -2.31	20	1.7	1	2	
-2.50 - -2.41	8	1.6	1	196	
-2.60 - -2.51	20	1.2	1	37	
-2.70 - -2.61	7	1.1	0		
-2.80 - -2.71	7	0.9	0		
-2.90 - -2.81	8	0.8	0		
-3.00 - -2.91	4	0.7	0		
-3.10 - -3.01	4	0.6	0		
-3.20 - -3.11	6	0.5	0		
-3.30 - -3.21	3	0.4	0		
-3.40 - -3.31	3	0.4	0		
-3.50 - -3.41	3	0.3	0		
-3.60 - -3.51	2	0.3	0		
-3.70 - -3.61	2	0.2	0		